

Combining pattern-based CRFs and weighted context-free grammars

Rustem Takhanov
takhanov@ist.ac.at

Vladimir Kolmogorov
vnk@ist.ac.at

Abstract

We consider two models for the sequence labeling (tagging) problem. The first one is a *Pattern-Based Conditional Random Field* (PB), in which the energy of a string (chain labeling) $x = x_1 \dots x_n \in D^n$ is a sum of terms over intervals $[i, j]$ where each term is non-zero only if the substring $x_i \dots x_j$ equals a prespecified word $w \in \Lambda$. The second model is a *Weighted Context-Free Grammar* (WCFG) frequently used for natural language processing. PB and WCFG encode local and non-local interactions respectively, and thus can be viewed as complementary.

We propose a *Grammatical Pattern-Based CRF model* (GPB) that combines the two in a natural way. We argue that it has certain advantages over existing approaches such as the *Hybrid model* of Benedí and Sanchez that combines *N-grams* and WCFGs. The focus of this paper is to analyze the complexity of inference tasks in a GPB such as computing MAP. We present a polynomial-time algorithm for general GPBs and a faster version for a special case that we call *Interaction Grammars*.

1 Introduction

The *sequence labeling* (or the *sequence tagging*) problem is a supervised learning problem with the following formulation: given an observation z (which is usually a sequence of n values), infer labeling $x = x_1 \dots x_n$ where each variable x_i takes values in some finite domain D . Such problem appears in many domains such as text and speech analysis, signal analysis, and bioinformatics. Standard approaches to this problem include Hidden Markov Models (*HMM*) and Conditional Random Fields (*CRFs*).

In many applications labelings x satisfy the following “sparsity” assumption: subwords $x_{i:j} \stackrel{\text{def}}{=} x_i \dots x_j$ of a fixed length $k = j - i + 1$ are distributed not uniformly over D^k , but are rather concentrated in a small subset of D^k . Words in this subset are called “patterns”; we will denote the set of patterns as $\Lambda \subseteq D^* = \bigcup_{k \geq 0} D^k$. Usually, Λ is taken as the set of short words (e.g. of length $k < 5$) that occur sufficiently often as subwords of labelings in the training data. For problems satisfying this assumption it is natural to define a model given by the probability distribution $p_\theta(x|z) = \frac{1}{Z} \exp\{-E_\theta^{pb}(x|z)\}$ with the energy function

$$E_\theta^{pb}(x|z) = \sum_{\alpha \in \Lambda} \sum_{\substack{[i,j] \subseteq [1,n] \\ j-i+1=|\alpha|}} \psi_{ij}^\alpha \cdot [x_{i:j} = \alpha] \quad (1)$$

where θ is a vector of parameters to be learned from data, ψ_{ij}^α is a function that can depend on z and θ , $|\alpha|$ is the length of word α and $[\cdot]$ is the *Iverson bracket* (i.e. $[s] = 1$ if statement s is true, otherwise $[s] = 0$). This model is called a *pattern-based CRF* (PB) [18, 16].

Intuitively, pattern-based CRFs allow to model long-range interactions that are carried through some selected sequences of labels. This could be useful in a variety of applications: in natural language processing patterns could correspond to certain syntactic constructions or stable idioms; in protein secondary structure prediction — to sequences of dihedral angles associated with stable configurations such as α -helices; in gene prediction — to sequences of nucleotides with supposed functional roles such as “exon” or “intron”, specific codons, etc.

But along with long-range interactions, modeled by a set of words Λ , there could be interactions that have a very non-local nature. Consider, for example, a language model problem, i.e. a problem of building a probabilistic model of sentences in a certain language. Standard and the simplest way to build a language model is *N-grams* approach, which is equivalent to representing the probability of a sentence as a pattern-based CRF where the set Λ is equal to the set of frequent *N-grams*. It is a well-known fact that for most of natural languages, a set of frequent *N-grams* for small N is not a large number, which justifies application of PB. But at the same time, sentences also have a syntactic structure that sometimes can be described by a context-free grammar. Such syntactic correlations could have a non-local structure, and cannot be encoded in the PB framework. Thus, we have a problem of introducing grammar-based structures into a model. One of approaches to implement this idea can be found, e.g., in [3].

Let us give another example of a problem that has a similar flavour. Suppose that our task is to identify, for a given RNA sequence $z \in \{G, A, U, C\}^n$, certain properties of each nucleotide z_i (e.g. discretized dihedral angles). Suppose we have 4 vocabularies of such properties D_G, D_A, D_U, D_C , one for each nucleotide. Thus, we have again the sequence labeling problem where labels alphabet is $D = \bigcup_t D_t$. It is natural to model local interactions by pattern-based models, if “sparsity” assumption is satisfied. But RNA’s secondary structure can also be modeled through context-free grammars (see Fig. 1). Now we can introduce a context-free grammar with nonterminals set $\{S\}$, terminals set V , and weighted rules $S \rightarrow SS$, $S \rightarrow a, a \in V$, $S \rightarrow xSy$, where $x \in D_t$, $y \in D_{t'}$ and t, t' are two complementary nucleotides. This is a generalization of the context-free grammar introduced in Fig. 1. Then any potential label-

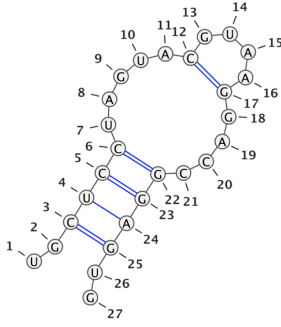


Figure 1: An RNA secondary structure without pseudo-knots. Every pair of complementary nucleotides that have hydrogen bonds with each other correspond to opening and closing brackets in a correct bracket structure:

UGCUCUAGUACGUAAGGACCGAGUG
 $\cdot \cdot ((((\cdot \cdot \cdot)))) \cdot \cdot \cdot$

Equivalently, the sequence can be parsed according to a context-free grammar defined by the set of nonterminals $\{S\}$, the set of terminals $\{G, A, U, C\}$, and rules $S \rightarrow SS$, $S \rightarrow G|A|U|C$, $S \rightarrow GSC|CSG|USA|ASU$.

ing for an input RNA sequence could be optimally parsed according to this grammar, and this parsing can be interpreted as a system of interacting nucleotides in the sequence, i.e. secondary structure. Moreover, the weight of the parsing is a sum of binary terms defined on labels where each term is an interaction potential between corresponding nucleotides. This weight could also be included into a probabilistic model. Note that in the last model an interaction that is associated with weighted rule $S \rightarrow xSy$ cannot be modeled by pattern-based CRFs.

These examples motivate the following model. Consider a weighted context-free grammar $\Gamma = (D, N, S, R, \nu)$, where N is the alphabet of nonterminals, $S \in N$ is the initial symbol, R is a set of rules, $\nu : R \rightarrow \mathbb{R}$ is a weighting function (that could depend on parameters θ and observation z). A *Grammatical Pattern-Based model* (GPB) is defined by probability distribution $p_\theta(x, \lambda|z) \sim \exp\{-E_\theta(x, \lambda|z)\}$ with the energy

$$E_\theta(x, \lambda|z) = E_\theta^{pb}(x|z) + C_{\Gamma(\theta)}(x, \lambda|z) \quad (2)$$

where $C_{\Gamma(\theta)}(x, \lambda|z)$ is the cost of derivation (parsing) λ of x according to $\Gamma(\theta)$. We view this as a rather natural way to combine PB and WCFG: defining energy as a sum of terms that encode different constraints has a long history in the CRF literature.

Contributions. This paper investigates the complexity of several inference tasks in a GPB. Our focus is on the problem of computing a Maximum a Posteriori (MAP) solution (x, λ) , i.e. minimizing energy (2). We show that this can be done in polynomial time; complexities are stated in the end of this section. We also discuss how our algorithms can be adapted to compute in polynomial time sums $\sum_{x, \lambda} \exp\{-E_\theta(x, \lambda|z)\}$ and $\sum_\lambda \exp\{-E_\theta(x, \lambda|z)\}$ (for given x, z).

Related work. Pattern-based CRFs with key inference algorithms first appeared in [18]. Refined versions of these

algorithms and an efficient sampling technique were described in [16]. Applications of PB considered in the literature so far include handwritten character recognition, identification of named entities from text [18], optical character recognition [14] and the protein dihedral angles prediction problem [16]. Further generalization of the model proposed in [14] considered a pattern as a set of strings rather than one single string. Another direction of generalization [12] extends the segments of variables on which patterns are defined by allowing correspondence of each label of a pattern to a successive repetition of it on a line.

Probably the closest to ours is a line of research represented by the work [3]. They proposed a probabilistic *Hybrid model* that also integrates local correlations (namely, the N -gram model which is a special case of PB) and stochastic grammars. Unlike us, they define the probability successively (their model is slightly different, but is equivalent to the following):

$$\begin{aligned} p(x_1 \dots x_n) &= \prod_{i=1}^n p(x_i | x_1 \dots x_{i-1}) \\ p(x_i | x_1 \dots x_{i-1}) &= \alpha p^{pb}(x_i | x_1 \dots x_{i-1}) + \\ &\quad + (1 - \alpha) p^\Gamma(x_i | x_1 \dots x_{i-1}) \\ p^\Gamma(x_i | x_1 \dots x_{i-1}) &= \frac{\sum_{y_{i+1:n}, \lambda} p^\Gamma(x_1 \dots x_i, y_{i+1:n}, \lambda)}{\sum_{y_{i:n}, \lambda} p^\Gamma(x_1 \dots x_{i-1}, y_{i:n}, \lambda)} \end{aligned}$$

where p^{pb} is an N -gram (a pattern-based) term and p^Γ is a grammar term (defined as $\sim e^{-C_\Gamma(x, \lambda)}$), and $\alpha \in [0, 1]$ is some parameter that mix two models. The Hybrid model have been applied to various problems, from modeling text to RNA secondary structure prediction [15].

Let us compare computational requirements of the two models, assuming that there is no dependence on observation z (as was the case in [3, 15]). Both GPB and the Hybrid models allow efficient computation of probability $p(x_1 \dots x_n)$ for a given word (for the former it equals $p(x) = \sum_\lambda p(x, \lambda)$). As can be shown, for GPB this can be done in $O(|R|n^3)$ time.¹ For the Hybrid model we would need $O(|R|n^3)$ for evaluating each term $p(x_i | x_1 \dots x_{i-1})$, resulting in $O(|R|n^4)$ total time.

Another important task is computing labeling x with the highest probability. A tricky definition of total probability in the Hybrid makes this a non-trivial problem (probably intractable), though sampling is easy. We conjecture that in GPB maximizing $p(x)$ over x is also intractable; however, we can do efficiently the next best thing, namely maximize $p(x, \lambda)$ over x, λ .

Therefore, we argue that GPB has computational advantages over the Hybrid model. Furthermore, a GPB model can be trained using standard techniques such as the maximum likelihood principle (with a gradient-based on an EM-like method) or the struct-SVM approach with hidden variables. The former requires computing sums of the form $\sum_{x, \lambda} \exp\{-E_\theta(x, \lambda)\}$ while the latter uses minimization of $E_\theta(x, \lambda)$ as a subroutine; both tasks can be performed in polynomial time.

¹We assume that the normalization constant $\sum_{x, \lambda} \exp\{-E_\theta(x, \lambda)\}$ has been precomputed; this can be done in polynomial time. Alternatively, this constant can be ignored if we are interested only in probability ratios.

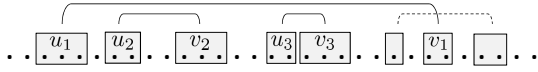


Figure 2: Interactions that can be modeled form a bracket structure. The interaction shown by a dashed line cannot be counted simultaneously with the interaction (u_1, v_1) .

Another way of mixing N -grams correlations with context-free grammars was given in [17]. Their model is quite different from ours, and uses an unweighted version of CFG.

Finally, we would like to mention that the idea of combining a certain sequential model with context-free grammars can be traced back to [2]. One of their classical results is as follows: if Γ is a context-free grammar given in Chomsky Normal Form (CNF) with m nonterminals and r rules and F is a finite-state automaton with s states, then intersection of languages defined by them is a context-free language that can be described by a context-free grammar with ms^2 nonterminals and rs^3 rules. With some work, this result together with a version of the CYK parsing algorithm from [11] could give an alternative way to derive the algorithm for general grammars. Instead of following this route, we chose to present the algorithm for general grammars directly.

Patterns interaction grammar. The algorithm for general WCFGs has a rather high complexity (although polynomial). An interesting question is whether there are special cases that admit faster inference. We identified one such case that we call *interaction grammars*; it is given by non-terminals set $N = \{S\}$ and a set of rules:

$$R = \left\{ \begin{array}{l} S \rightarrow SS; S \rightarrow a \in D \cup \{\varepsilon\} \\ S \rightarrow uSv, (u, v) \in P \end{array} \right\} \quad (3)$$

where P is a certain subset of $\Lambda \times \Lambda$. We will denote such grammar as $\Gamma(P)$. We also restrict that only rules of the third type can have nonzero weights. Note that the grammar that we described in the second motivating problem example, i.e. RNA sequence labeling, is of this kind.

Interaction grammars strengthen the PB model by allowing non-local interactions between patterns, albeit with some limitations on such interactions. Roughly speaking, two interactions (u, v) and (u', v') can be counted simultaneously only if they are either nested or do not overlap (Fig. 2).

For computational reasons we will also consider a further restriction in which the depth of inclusion does not exceed some constant d . (In the example in Fig. 2 this depth equals 2.) Such restriction can be expressed by the grammar with non-terminals $N = \{S^0, \dots, S^{d-1}, S = S^d\}$ and with the following set of rules:

$$R = \left\{ \begin{array}{l} S^k \rightarrow S^k S^k, k = 0, \dots, d \\ S^k \rightarrow S^{k-1}; S^0 \rightarrow a \in D \cup \{\varepsilon\} \\ S^k \rightarrow uS^{k-1}v, k = 1, \dots, d, (u, v) \in P \end{array} \right\} \quad (4)$$

Again, only rules of the fourth type can have nonzero weights. We call it an interaction grammar of depth d .

Minimization. This paper focuses on minimization algorithms for energies of the form (2) over λ, x . Without loss of generality we can eliminate λ by defining new functional that depends only on x :

$$E_w(x|z) = E_\theta^{pb}(x|z) + C_\Gamma(x|z) \quad (5)$$

where $C_\Gamma(x|z)$ is the cost of a least-weight derivation of x according to Γ .

We will consider the following three cases: (i) general WCFG; (ii) interaction grammar of depth $d \geq 2$; (iii) interaction grammar of depth 1. For all three cases we will present algorithms. The complexity of solving these tasks is discussed below. We denote $L = \sum_{\alpha \in \Gamma} |\alpha|$ to be total length of patterns and $\ell_{\max} = \max_{\alpha \in \Gamma} |\alpha|$ to be the maximum length of a pattern.

For the most general case (i) we present $\Theta(|R|(nL)^3)$ algorithm that uses $\Theta(|N|(nL)^2)$ space if grammar is given in CNF. This algorithm is based on dynamic programming and uses a very similar data structures (so called *messages*) to standard CYK parsing algorithm [1]. Thus, a standard way of obtaining CNF leads us to $\Theta((|P| + |D|)(nL)^3)$ algorithm for interaction grammars and $\Theta(d(|P| + |D|)(nL)^3)$ for interaction grammars of depth d . Note that at the core of the algorithm we compute multiplication of two $nL \times nL$ matrices over a semiring $(S, \oplus, \otimes) = (\mathbb{R}, \max, +)$ (so called *min-plus product*, or *the distance product*) which makes it cubic with respect to nL . It is well-known [20] that the distance product of matrices is computationally equivalent to *all-pairs shortest path problem*, and this leads us to more efficient algorithms for special cases of WCFGs, namely interaction grammars of depth d .

In this case we compute a similar set of messages that are defined on $d + 1$ levels (from 0 to d). There are two types of operations that we apply to messages at d iterations: first we compute messages of level i based on already computed messages of level $i - 1$ (we call this *vertical messages passing*), second we solve the all-pairs shortest path problem on a certain graph. In the worst case, total complexity of this algorithm for interaction grammars of depth d does not differ from complexity of the general algorithm. However, in the best case it can be much faster, namely $O((nL)^2(d|P| + d \log nL + |D|))$. Computational results on some synthetic data are given in section 6. Moreover, when $d = 1$, the complexity is always $O(|P|nL(\ell_{\min} \min(|D|, \log \ell_{\min}) + |P|))$ where $\ell_{\min} = \min_{w \in \Lambda} |w|$.

2 Notation and preliminaries

First, we introduce a few definitions.

- A *pattern* is a pair $\alpha = ([i, j], w)$ where $[i, j]$ is an interval in $[1, n]$ and $w = w_i \dots w_j$ is a sequence over alphabet D indexed by integers in $[i, j]$ ($j \geq i - 1$). The *length* of α is denoted as $|\alpha| = |w| = j - i + 1$. For pattern $\alpha = ([i, j], w)$ we will also denote $i_\alpha = i, j_\alpha = j, w_\alpha = w$.
- Symbols “*” denotes an arbitrary word or pattern (possibly the empty word ε or the empty pattern $\varepsilon_s \triangleq$

$([s+1, s], \varepsilon)$ at position s). The exact meaning will always be clear from the context. Similarly, “+” denotes an arbitrary non-empty word or pattern.

- The concatenation of patterns $\alpha = ([i, j], v)$ and $\beta = ([j+1, k], w)$ is the pattern $\alpha\beta \triangleq ([i, k], vw)$. Whenever we write $\alpha\beta$ we assume that it is defined, i.e. $\alpha = ([\cdot, j], \cdot)$ and $\beta = ([j+1, \cdot], \cdot)$ for some j . Also, if $u \in D^*$, then $\alpha u = ([i, j+|u|], vu)$ and $u\alpha = ([i-|u|, j], uv)$.
- For a pattern $\alpha = ([i, j], v)$ and interval $[k, \ell] \subseteq [i, j]$, the *subpattern of α at position $[k, \ell]$* is the pattern $\alpha_{k:\ell} \triangleq ([k, \ell], v_{k:\ell})$ where $v_{k:\ell} = v_k \dots v_\ell$. If $k = i$ then $\alpha_{k:\ell}$ is called a *prefix* of α . If $\ell = j$ then $\alpha_{k:\ell}$ is a *suffix* of α .
- If β is a subpattern of α , i.e. $\beta = \alpha_{k:\ell}$ for some $[k, \ell]$, then we say that β is *contained* in α . This is equivalent to the condition $\alpha = *\beta*$.
- $D^{i:j} = \{([i, j], v) \mid v \in D^{[i, j]}\}$ is the set of patterns with interval $[i, j]$.
- For a pattern α let α^- be the prefix of α of length $|\alpha|-1$; if α is empty then α^- is undefined.

We will consider the following problem. Let Π° be the set of patterns of words in Λ placed at all possible positions: $\Pi^\circ = \{([i, j], \alpha) \mid \alpha \in \Lambda\}$. Define the cost of pattern $x \in D^{i:j}$ via

$$f(x) = \sum_{\alpha \in \Pi^\circ, x = *\alpha*} c_\alpha \quad (6)$$

where $c_\alpha \in R, \alpha \in \Pi^\circ$ are fixed constants. Let $C_\Gamma(x)$ be the cost of a least-weight derivation of x in Γ . Our goal is to compute

$$M = \min_{x \in D^{1:n}} F(x) \quad (7)$$

where $F(x) = f(x) + C_\Gamma(x)$.

We select set Π as the set of prefixes of patterns in Π° :

$$\Pi = \{\alpha \mid \exists \alpha* \in \Pi^\circ\} \quad (8)$$

- For an index $s \in [0, n]$ we denote Π_s to be the set of patterns in Π that end at position s : $\Pi_s = \{\alpha \in \Pi \mid j_\alpha = s\}$.
- For an arbitrary pattern α , $lsp(\alpha)$ denotes the longest suffix of α that belongs to Π_{j_α} .

Graph $G[\Pi_s]$. The following construction will be used throughout the paper. Given a set of patterns Π and index s , we define $G[\Pi_s] = (\Pi_s, E[\Pi_s])$ to be a directed graph with the following set of edges: (α, β) belongs to $E[\Pi_s]$ for $\alpha, \beta \in \Pi_s$ if α is a proper suffix of β ($\beta = +\alpha$) and Π_s does not have an “intermediate” suffix γ of β with $|\beta| > |\gamma| > |\alpha|$. It can be checked that graph $G[\Pi_s]$ is a directed forest. If $\varepsilon_s \in \Pi_s$ then $G[\Pi_s]$ is connected and therefore is a tree. In this case we treat ε_s as the root. An example is shown in Fig. 3.

Computing partial costs. Recall that $f(\alpha)$ for a pattern α is the cost of all patterns inside α (eq. (6)). We also define $\phi(\alpha)$ to be the cost of only those patterns that are suffixes of α :

$$\phi(\alpha) = \sum_{\beta \in \Pi^\circ, \alpha = *\beta*} c_\beta \quad (9)$$

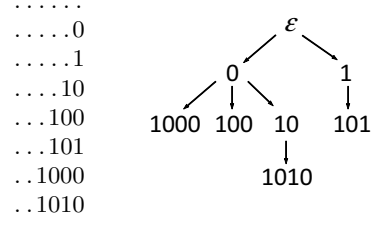


Figure 3: Graph $G[\Pi_s]$ for the set of 8 patterns shown on the left (for brevity, their intervals are not shown; they all end at the same position s .) This set of patterns would arise if $\Gamma = \{0, 1, 1000, 1010\}$ and Π was defined as the set of all prefixes of patterns in Π° .

In algorithms below we will use the following quantities: $\phi(\alpha), f(\alpha)$ for $\alpha \in \Pi$ and $f(\alpha\beta)$ for $\alpha, \beta \in \Pi$. Let us show how to compute them efficiently.

Lemma 1 *Values $\phi(\alpha), f(\alpha)$ for all $\alpha \in \Pi$ can be computed using $O(|\Pi|)$ additions and values $f(\alpha\beta), \alpha, \beta \in \Pi$ can be computed using $O(L|\Pi|)$ additions.*

Proof. To compute $\phi(\cdot)$ for patterns $\alpha \in \Pi_s$, we use the following procedure: (i) set $\phi(\varepsilon_s) := 1$; (ii) traverse edges $(\alpha, \beta) \in E[\Pi_s]$ of tree $G[\Pi_s]$ (from the root to the leaves) and set

$$\phi(\beta) := \begin{cases} \phi(\alpha) + c_\beta & \text{if } \beta \in \Pi^\circ \\ \phi(\alpha) & \text{otherwise} \end{cases}$$

After computing $\phi(\cdot)$, we go through indexes $s \in [0, n]$ in increasing order and set

$$f(\varepsilon_s) := 0, \quad f(\alpha) := f(\alpha^-) + \phi(\alpha) \quad \forall \alpha \in \Pi_s - \{\varepsilon_s\}$$

$$f(\alpha\beta) := f(\alpha\beta^-) + \phi(lsp(\alpha\beta)), \quad \forall \alpha \in \Pi_{i_\beta-1}, \beta \in \Pi_s - \{\varepsilon_s\}$$

□

3 Minimization for general grammars

In this section we describe our algorithm for minimizing energy (7) for general grammars. The idea is to compute certain “messages” by a dynamic programming procedure. Note that the same system of messages will be computed by algorithms for interaction grammars, though the computation strategies will differ.

We will assume that a WCFG $\Gamma = (D, N, S, R, \nu)$ is given in CNF, i.e. all of its rules belong to one of the following types:

$$\begin{aligned} A &\rightarrow BC, \quad A, B, C \in N \\ A &\rightarrow w, \quad A \in N, w \in D^* \\ S &\rightarrow \varepsilon \end{aligned}$$

We will also assume that for all rules $A \rightarrow w \in R$, the word w belongs to Λ . The set Π has a natural partial order \succ on it, such that for $\alpha = ([i_\alpha, j_\alpha], w_\alpha)$ and $\beta = ([i_\beta, j_\beta], w_\beta)$:

$$\beta \succ \alpha \Leftrightarrow j_\beta > j_\alpha, i_\beta \geq i_\alpha, \alpha_{i_\beta:j_\alpha} = \beta_{i_\beta:j_\alpha} \quad (10)$$

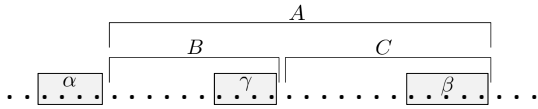


Figure 4: Computing message $M_A(\alpha, \beta)$ assuming that $A \rightarrow BC$ is the next rule to be applied and γ is the longest suffix for B which is in Π .

In the previous definition if $i_\beta > j_\alpha$, then the condition $\alpha_{i_\beta:j_\alpha} = \beta_{i_\beta:j_\alpha}$ is satisfied by definition.

We will compute messages $M_A(\alpha, \beta)$ for $A \in N$ and any pair of patterns $\alpha, \beta \in \Pi$ such that $\beta \succ \alpha$. Message $M_A(\alpha, \beta)$ has the following interpretation. We consider interval of variables $x_{i_\alpha:j_\beta}$ and all their assignments such that $x_{i_\alpha:j_\alpha} = w_\alpha$, $x_{i_\beta:j_\beta} = w_\beta$ and for any $(\beta, \gamma) \in E[\Pi_{j_\beta}]$, $x_{i_\gamma:j_\gamma} \neq w_\gamma$. On these variables we form a new functional that consists of those patterns (of PB part) of (7) that belong to this interval plus $C_{\Gamma_A}(x_{j_\alpha+1:j_\beta})$ where Γ_A is the same as Γ but with the initial symbol changed to A . A minimum of this functional over defined set of assignments is our message $M_A(\alpha, \beta)$. I.e.:

$$M_A(\alpha, \beta) = \min_{\substack{x: x=\alpha * = \beta \\ \forall (\beta, \gamma) \in E[\Pi_{j_\beta}]: \\ x \neq * \gamma}} f(x) + C_{\Gamma_A}(x_{j_\alpha+1:j_\beta}) \quad (11)$$

As in the CYK algorithm, the main idea of our approach is to compute message $M_A(\alpha, \beta)$ based on the assumption that the best parsing of $x_{j_\alpha+1:j_\beta}$ for optimal x starts by applying the rule $A \rightarrow BC$. Therefore, we need to consider possible divisions of $[j_\alpha + 1, j_\beta]$ into two parts corresponding to nonterminals B and C . To count patterns that cross the boundary between B and C , we also need to know the pattern γ with which the first part ends (Fig. 4); such “crossing” patterns will be included in $M_C(\gamma, \beta)$. The resulting procedure is given in Algorithm 1. Note that in (13) we subtract $f(\gamma)$ to avoid counting patterns inside γ twice.

Algorithm 1 Computing minimum of $F(x)$

- 1: set $M_A(\alpha, \beta) := +\infty$ for all messages
- 2: **for** each rule $A \rightarrow w \in R$ and each $\alpha \in \Pi$ set

$$M_A(\alpha, \text{fsp}(\alpha w)) := f(\alpha w) + \nu(A \rightarrow w) \quad (12)$$

- 3: **for** each $\alpha, \beta \in \Pi : \beta \succ \alpha$ in the order of increasing $j_\beta - j_\alpha$ set $M_A(\alpha, \beta) :=$

$$\min_{r=A \rightarrow BC \in R} \min_{\substack{\gamma \in \Pi: \\ \beta \succ \gamma \succ \alpha}} M_B(\alpha, \gamma) + M_C(\gamma, \beta) - f(\gamma) + \nu(r) \quad (13)$$

- 4: **return** $M := \min_{\alpha \in \Pi_n} M_S(\varepsilon_0, \alpha)$
-

Theorem 2 Algorithm 1 is correct, i.e. it returns the correct value of $M = \min_x F(x)$.

3.1 Proof of Theorem 2

First, we will prove the following result.

Lemma 3 For each $\alpha, \beta, \gamma \in \Pi : \beta \succ \gamma \succ \alpha$ and a rule $r = A \rightarrow BC \in R$:

$$M_A(\alpha, \beta) \leq M_B(\alpha, \gamma) + M_C(\gamma, \beta) - f(\gamma) + \nu(r)$$

Proof. Suppose that $x_{i_\alpha:j_\gamma}^1$ is an optimal argument for message $M_B(\alpha, \gamma)$ and $x_{i_\gamma:j_\beta}^2$ is an optimal argument for message $M_C(\gamma, \beta)$. Since both these arguments are equal to w_γ on their intersection, there is $x_{i_\alpha:j_\beta}$ such that $x_{i_\alpha:j_\gamma}^1 = x_{i_\alpha:j_\gamma}$ and $x_{i_\gamma:j_\beta}^2 = x_{i_\gamma:j_\beta}$. Then if we use $x_{i_\alpha:j_\beta}$ as an argument for message $M_A(\alpha, \beta)$ and parse $x_{j_\alpha+1:j_\beta}$ by applying rule $A \rightarrow BC$ and then deriving $x_{j_\alpha+1:j_\gamma}$ from B and $x_{j_\gamma+1:j_\beta}$ from C (optimally, i.e. with minimal weight), we obtain an upper bound for $M_A(\alpha, \beta)$ that equals $M_B(\alpha, \gamma) + M_C(\gamma, \beta) - f(\gamma) + \nu(r)$. Here we needed to subtract $f(\gamma)$ in order to avoid counting some patterns twice. \square

Let us now prove by induction that Algorithm 1 correctly computes all messages. It can be checked that in step 2, the statement that $\beta = \text{fsp}(\alpha w)$, is equivalent that $\beta \in \Pi$ is such that $\beta \succ \alpha$, $j_\beta - j_\alpha = |w|$, $\alpha w = * \beta$ and $\alpha w \neq \gamma$ for any $(\beta, \gamma) \in E[\Pi_{j_\beta}]$. Or, equivalently, $x = \alpha w$ can serve as an argument for message $M_A(\alpha, \beta)$ (see (11) for the definition of the set of arguments).

Then we simply compute messages $M_A(\alpha, \beta)$ based on the assumption that $f(x) + C_{\Gamma_A}(x_{j_\alpha+1:j_\beta})$ (see message definition) attains its minimum when we set $x = \alpha w$ and $x_{j_\alpha+1:j_\beta} = w$ is parsed via rule $A \rightarrow w$. In this case, $C_{\Gamma_A}(x_{j_\alpha+1:j_\beta}) = \nu(A \rightarrow w)$ and $f(x) = f(\alpha w)$ which explains the expression to be computed.

Those messages for which we computed their values correctly in steps 1-2 will serve as an induction base. Now let us consider a message $M_A(\alpha, \beta)$ for which steps 1-2 compute non-optimal value. This means that an optimal argument $x_{j_\alpha+1:j_\beta}$ should be parsed by applying first some rule $A \rightarrow BC$. Suppose that in an optimal parsing nonterminal B is responsible for a word $x_{j_\alpha+1:t}$ and nonterminal C for residual $x_{t+1:j_\beta}$. Consider all patterns that are satisfied on optimal x and contain variable x_t . Suppose that γ is one of them which has the leftmost i_γ . Denote $\gamma' = \gamma_{i_\gamma:t}$. Then $x_{i_\alpha:t}$ is also an optimal argument for message $M_B(\alpha, \gamma')$ and $x_{t+1-|\gamma'|:j_\beta}$ is an optimal argument for $M_C(\gamma', \beta)$. It is easy to check that, otherwise, we could change x in segment $x_{i_\alpha:t}$ (or in segment $x_{t+1-|\gamma'|:j_\beta}$) to a locally optimal that will improve an overall value. Therefore,

$$M_A(\alpha, \beta) = M_B(\alpha, \gamma') + M_C(\gamma', \beta) - f(\gamma') + \nu(A \rightarrow BC)$$

Together with Lemma 3 this implies that $M_A(\alpha, \beta)$ is equal to the right side of (13).

Now we only need to notice that we reduced the computation of a message to other messages with smaller $j_\beta - j_\alpha$. Therefore, correctness of the computation depends on whether we calculate correctly messages $M_A(\alpha, \beta)$ with such small values of $j_\beta - j_\alpha$, that a part of their optimal argument $x_{j_\alpha+1:j_\beta}$ cannot be parsed by a rule of the form $A \rightarrow BC$ (i.e. cannot be split). But, since we already know that we computed such messages correctly, we conclude that all final message values are correct.

3.2 Algorithm's complexity

The complexity of step 1 is $O((nL)^2)$. In step 2 we go through at most $nL|R|$ different pairs $(A \rightarrow w, \alpha)$: $|R|$ for $A \rightarrow w$, nL for α . Since w_β is a function of w_α and w , it can be computed during preprocessing. It can be seen that with proper preprocessing of the set of pairs $\{(A \rightarrow w, w_\alpha)\}$ we can compute $M_A(\alpha, \beta)$ in step 2 in time $O(1)$. Given that we already computed $f(\alpha w)$ according to Lemma 1, the complexity of the step 2 is bounded by $O(nL|R|)$.

Finally, the complexity of step 3 can be bounded by $O(|R||\Pi|^3) = O(|R| \cdot (nL)^3)$ which clearly dominates two previous steps and the computation of values in Lemma 1. Therefore, the algorithm's overall complexity is $O(|R|(nL)^3)$.

4 Algorithm for interaction grammars

We will now describe an algorithm for interaction grammars of depth d , $\Gamma^d(P)$. Recall that only rules of the form $S^k \rightarrow uS^{k-1}v$, $(u, v) \in P$ can have nonzero weight; we denote this weight as $\theta_{u,v}^k$. Unlike the general grammar case, we will not convert the grammar into a CNF, but will use the initial form of it. Our algorithm is based on computing the same set of messages $M_{S^k}(\alpha, \beta)$, $k = 0, \dots, d$, which we will now denote simpler as $M_k(\alpha, \beta)$.

We will successively compute messages $M_k(\alpha, \beta)$ for $k = 0, 1, \dots, d$. At the k th iteration we first compute message $M_k(\alpha, \beta)$ based on the assumption that optimal parsing of $x_{j_\alpha+1:j_\beta}$ starts by applying some rule $S^k \rightarrow uS^{k-1}v$ (step 3 of Algorithm 2 below). This part of computation can be done relatively fast. Then we have to update already computed messages $M_k(\alpha, \beta)$ based on the assumption that optimal parsing starts by dividing the argument $x_{j_\alpha+1:j_\beta}$ into pieces (by applying the rule $S^k \rightarrow S^k S^k$) (step 4). As we will see, this is equivalent to computing shortest paths in a graph with vertex set Π .

Theorem 4 *Algorithm 2 is correct, i.e. it returns the correct value of $M = \min_x F(x)$ for an interaction grammar of depth d .*

4.1 Proof of Theorem 4

Lemma 5 *Suppose that $\alpha, \beta \in \Pi : \beta \succ \alpha$ and optimal argument $x_{i_\alpha:j_\beta}$ for message $M_k(\alpha, \beta)$ are such that optimal parsing of $x_{j_\alpha+1:j_\beta}$ from nonterminal S^k starts by applying the rule $S^k \rightarrow uS^{k-1}v$. Then,*

$$M_k(\alpha, \beta) = \min_{\gamma \in \Omega} M_{k-1}(lsp(\alpha u), \gamma) + \theta_{u,v}^k + f(\alpha u) - f(lsp(\alpha u)) + f(\gamma v) - f(\gamma) \quad (16)$$

where

$$\Omega = \left\{ \begin{array}{l} \gamma \in \Pi | \gamma = * \beta_{1:|\beta|-|v|}, \forall (\beta, \delta) \in E[\Pi_{j_\beta}], \\ \gamma \neq * \delta_{1:|\delta|-|v|} \end{array} \right\}$$

Algorithm 2 Computing minimum of $F(x)$

- 1: **for** each $\alpha, \beta \in \Pi : \beta \succ \alpha$ compute $M_0(\alpha, \beta)$
- 2: **for** $k = 1, \dots, d$ **do**
- 3: **for** each $\alpha, \beta \in \Pi : \beta \succ \alpha$

$$M_k(\alpha, \beta) := \min_{(u,v,\gamma) \in \Omega(\beta)} M_{k-1}(lsp(\alpha u), \gamma) + \theta_{u,v}^k + f(\alpha u) - f(lsp(\alpha u)) + f(\gamma v) - f(\gamma) \quad (14)$$

where $\Omega(\beta) = \{(u, v, \gamma) | (u, v) \in P, x_\beta = *v; \gamma \in \Pi, \gamma = * \beta_{1:|\beta|-|v|}, \forall (\beta, \delta) \in E[\Pi_{j_\beta}] \gamma \neq * \delta_{1:|\delta|-|v|}\}$.

- 4: **compute** all-pairs shortest paths (denoted $SP(\alpha, \beta)$) for oriented graph $G = (\Pi, \succ)$ with costs of edges $c(\alpha, \beta) := M_k(\alpha, \beta) - f(\beta)$
- 5: **for** each $\alpha, \beta \in \Pi : \beta \succ \alpha$

$$M_k(\alpha, \beta) := SP(\alpha, \beta) + f(\beta) \quad (15)$$

6: **end for**

7: **return** $M := \min_{k=0,d} \min_{\alpha \in \Pi_n} M_k(\varepsilon_0, \alpha)$

Proof. Since first rule in the parsing of $x_{j_\alpha+1:j_\beta}$ is $S^k \rightarrow uS^{k-1}v$, then $x_{j_\alpha+1:j_\beta} = *v$. Therefore, $w_\beta = *v$ or $+w_\beta = v$. The second option does not hold because we require that $\forall (\beta, \delta) \in E[\Pi_{j_\beta}] x_{i_\alpha:j_\beta} \neq * \delta$.

Suppose now that γ is the longest pattern from $\Pi_{j_\beta-|v|}$ for which $x_{i_\alpha:j_\beta-|v|} = * \gamma$. According to the definition, $M_k(\alpha, \beta) = f(x_{i_\alpha:j_\beta}) + C_{\Gamma_{S^k}}(x_{j_\alpha+1:j_\beta})$. We know that $C_{\Gamma_{S^k}}(x_{j_\alpha+1:j_\beta}) = \theta_{u,v}^k + C_{\Gamma_{S^{k-1}}}(x_{j_\alpha+1+|u|:j_\beta-|v|})$. Now we will prove that

$$f(x_{i_\alpha:j_\beta}) = f(x_{j_\alpha+1+|u|:j_\beta-|v|}) + f(\alpha u) + f(\gamma v) - f(lsp(\alpha u)) - f(\gamma) \quad (17)$$

First check that if a pattern $\delta \in \Pi_0$ is contained in $x_{i_\alpha:j_\beta}$ and is not counted in the first term of (17), then we have only 2 options: either δ is contained in 1) αu , or in 2) γv . Indeed, δ is not inside $[j_\alpha+1+|u|:j_\beta-|v|]$ only if either a) $j_\delta < j_\alpha+1+|u|-|lsp(\alpha u)|$ (then, obviously, δ is in αu), or b) $i_\delta > j_\beta-|v|$ (then δ is in γv), or c) $j_\alpha+|u|-|lsp(\alpha u)|, j_\alpha+1+|u|-|lsp(\alpha u)| \in [i_\delta, j_\delta]$, or d) $j_\beta-|v|, j_\beta-|v|+1 \in [i_\delta, j_\delta]$. In case c) δ is not in αu only if $j_\delta > j_\alpha+|u|$ which implies that $* \delta_{i_\delta:j_\alpha+|u|} = \alpha u, \delta_{i_\delta:j_\alpha+|u|} \in \Pi$, and therefore $* \delta_{i_\delta:j_\alpha+|u|} = lsp(\alpha u)$, which contradicts that δ contains $x_{j_\alpha+1+|u|:j_\beta-|v|}$. In case d) $\delta_{i_\delta:j_\beta-|v|} \in \Pi$ and by definition of γ , $\gamma = * \delta_{i_\delta:j_\beta-|v|}$, which implies that δ is in γv .

Formula (17) is obtained from inclusion-exclusion principle. Indeed, patterns in $x_{i_\alpha:j_\beta}$, as we proved, is a union of patterns in $x_{j_\alpha+1+|u|:j_\beta-|v|}$, αu and γv . Moreover, intersection of a set of patterns in $x_{j_\alpha+1+|u|:j_\beta-|v|}$ and a set of patterns in αu (γv) is a set of patterns in $lsp(\alpha u)$ (γ). The last thing we need to check is that patterns that are in both αu and in γv are all in $x_{j_\alpha+1+|u|:j_\beta-|v|}$. It is easy to see that this is equivalent to $[i_\gamma, j_\alpha+|u|] \subseteq [j_\alpha+1+|u|-|lsp(\alpha u)|, j_\beta-|v|]$ (if the first interval is empty then statement is obvious). The fact that $j_\alpha+|u| < j_\beta-|v|$ is obvious. Since $\gamma_{i_\gamma:j_\alpha+|u|} \in \Pi$ and

$\alpha u = * \gamma_{i_\gamma:j_\alpha+|u|}$, then by definition of $lsp(\alpha u)$ we conclude that $i_\gamma > j_\alpha + |u| - lsp(\alpha u)$. Equality (17) is proved.

Now we see that

$$M_k(\alpha, \beta) = f(x_{j_\alpha+1+|u|-|lsp(\alpha u)|:j_\beta-|v|}) + C_{\Gamma_{S^{k-1}}}(x_{j_\alpha+1+|u|:j_\beta-|v|}) + \theta_{u,v}^k + f(\alpha u) + f(\gamma v) - f(lsp(\alpha u)) - f(\gamma) \quad (18)$$

Note that $x_{j_\alpha+1+|u|-|lsp(\alpha u)|:j_\beta-|v|}$ can serve as an argument for message $M_{k-1}(lsp(\alpha u), \gamma)$. Indeed, by definition of γ , for any $(\gamma, \delta) \in \Pi_{\gamma, x_{j_\alpha+1+|u|-|lsp(\alpha u)|:j_\beta-|v|}} \neq * \delta$ and we can consider any parsing of $x_{j_\alpha+1+|u|:j_\beta-|v|}$ from S^{k-1} . It can be shown that $f(x_{j_\alpha+1+|u|-|lsp(\alpha u)|:j_\beta-|v|}) + C_{\Gamma_{S^{k-1}}}(x_{j_\alpha+1+|u|:j_\beta-|v|})$ is equal to $M_{k-1}(lsp(\alpha u), \gamma)$, since formula (17) holds for any variations of x for which α, β, γ preserve their properties. Therefore we conclude that $M_k(\alpha, \beta) = M_{k-1}(lsp(\alpha u), \gamma) + \theta_{u,v}^k + f(\alpha u) - f(lsp(\alpha u)) + f(\gamma v) - f(\gamma)$. To obtain the final formula we need to add minimum over all γ . \square

Lemma 6 Let $G = (\Pi, \succ)$ be a weighted oriented graph with costs of edges $c(\alpha, \beta) = \widetilde{M}_k(\alpha, \beta) - f(\beta)$, where $\widetilde{M}_k(\alpha, \beta)$ are values of messages after step 3 of Algorithm 2. Suppose that $\alpha, \beta \in \Pi : \beta \succ \alpha$. Then, $M_k(\alpha, \beta) = SP(\alpha, \beta) + f(\beta)$, where $SP(\alpha, \beta)$ is the length of a shortest path from α to β in graph G .

Proof. Suppose that optimal argument $x_{i_\alpha:j_\beta}$ for message $M_k(\alpha, \beta)$ is such that optimal parsing of $x_{j_\alpha+1:j_\beta}$ from nonterminal S^k starts by applying the rule $S^k \rightarrow S^k S^k$, then on one of resulting S^k we again apply the same rule and etc.: $S^k \rightarrow S^k S^k \rightarrow S^k S^k S^k \rightarrow \dots \rightarrow \underbrace{S^k \dots S^k}_{s \text{ times}}$,

and after this each S^k unfolds according to some $S^k \rightarrow u S^{k-1} v \in R$.

Suppose that $x_{j_\alpha+1:l_1}$ is parsed from the first S^k , $x_{l_1+1:l_2}$ is parsed from the second S^k and etc., until $x_{l_{s-1}+1:l_s}, l_s = j_\beta$, is parsed from the last S^k . Also, γ_i is longest pattern in Π_{l_i} for which $x_{i_\alpha:l_i} = * \gamma_i$. Clearly, $\gamma_s = \beta$. For completeness, $\gamma_0 = \alpha$. Then it is easy to check that $\gamma_i \succ \gamma_{i-1}$ and (we give it without proof)

$$M_k(\alpha, \beta) - f(\beta) = \sum_{i=1}^s \widetilde{M}_k(\gamma_{i-1}, \gamma_i) - f(\gamma_i) \quad (19)$$

We subtract $f(\gamma_i)$ each time in order avoid counting some patterns twice (like it would be in $\widetilde{M}_k(\gamma_{i-1}, \gamma_i) + \widetilde{M}_k(\gamma_i, \gamma_{i+1})$).

And visa versa, if we have a chain $\gamma_s \succ \gamma_{s-1} \succ \dots \succ \gamma_0$, then it corresponds to some pattern $x_{i_{\gamma_0}:j_{\gamma_s}}$ that is defined by requirement that $x_{i_{\gamma_i}:j_{\gamma_{i+1}}}$ is an optimal argument for message $\widetilde{M}_k(\gamma_i, \gamma_{i+1})$. A parsing of $x_{j_{\gamma_0}+1:j_{\gamma_s}}$ consists of first applying $s-1$ times the rule $S^k \rightarrow S^k S^k$, and deriving $x_{j_{\gamma_{i-1}}+1:j_{\gamma_i}}$ with minimal weight from i -th S^k . The cost of such parsing plus pattern-based part is equal to $f(\gamma_i) + \sum_{i=1}^s \widetilde{M}_k(\gamma_{i-1}, \gamma_i) - f(\gamma_i)$.

Now we see that an optimal argument with its parsing corresponds to a sum of the form (19) and each such sum

corresponds to some argument and its parsing. Therefore, an optimal argument with its parsing for $M_k(\alpha, \beta)$ can be found by computing a shortest path from α to β in G and the value of message is defined by (15). \square

Lemmas 5 and 6 imply that in step 5 of Algorithm 2 we correctly compute messages $M_k(\alpha, \beta)$. It remains to verify that the expression in step 7 gives indeed the value of the optimum M ; this fact follows directly from definitions.

4.2 Analysis of complexity

First let us estimate the complexity of step 1. A message $M_0(\alpha, \beta)$ does not include any grammatical part and can be computed with the same techniques as in [16]. There it was proved that all messages (in current notations) of the form $M_0(\varepsilon_0, \beta)$ for all $\beta \in \Pi$ can be computed in time $O(nL|D|)$. Now we only have to note that any $\alpha \in \Pi$ defines its own pattern-based potential on the interval of variables $[i_\alpha, n]$ that includes all patterns from Π° that satisfy: a) a pattern interval is a subinterval of $[i_\alpha, n]$; b) a pattern is consistent with α on the intersection with $[i_\alpha, j_\alpha]$. We assume weights of patterns to be the same, except for $c_\alpha = -C + c_\alpha^{old}$ where C is a large constant. For each such pattern-based potential p_α we can compute in $O(nL|D|)$ time all messages of the form $M_0^{p_\alpha}(\varepsilon_{i_\alpha-1}, \beta)$ that will be equal to $M_0(\alpha, \beta) - C$. Therefore, all messages $M_0(\alpha, \beta)$ can be computed in $O((nL)^2|D|)$ time.

Now let us turn to step 3 of the algorithm where we make vertical message passing. The complexity of this part is $O\left(dnL \sum_{\beta \in \Pi} |\Omega(\beta)|\right)$: d times in a loop, $O(nL)$ variants of choosing α , and $|\Omega(\beta)|$ variants for choosing (u, v, γ) given $\beta \in \Pi$.

Lemma 7 $\sum_{\beta \in \Pi} |\Omega(\beta)| \leq |P|nL$.

Proof. Let us consider set $X = \{(\beta, u, v, \gamma) | \beta \in \Pi, (u, v, \gamma) \in \Omega(\beta)\}$. The cardinality of this set exactly equals the expression to be bounded. Now given $(u, v) \in P$ and $\gamma \in \Pi$ we will show that if $(\beta, u, v, \gamma) \in X$ then $\beta = lsp(\gamma v)$; this will imply the lemma.

Using definition of $\Omega(\beta)$ we reformulate X as $\{(\beta, u, v, \gamma) | \beta \in \Pi, (u, v) \in P, x_\beta = *v; \gamma \in \Pi, \gamma = * \beta_{1:|\beta|-|v|}, \forall (\beta, \delta) \in E[\Pi_{j_\beta}] \gamma \neq * \delta_{1:|\delta|-|v|}\}$. Then with fixed $(u, v) \in P$ and $\gamma \in \Pi$ the definition requires that β satisfies: $\beta \in \Pi, x_\beta = *v, \gamma v = * \beta$ and $\forall (\beta, \delta) \in E[\Pi_{j_\beta}] \gamma v \neq * \delta$. It can be checked that this is equivalent to $\beta = lsp(\gamma v)$. \square

From this lemma we conclude that complexity of step 3 is $O(d(nL)^2|P|)$.

Step 4 is the most expensive step of the algorithm. To estimate its complexity, we need to specify how we compute shortest paths.

Theorem 8 $M = \min_x F(x)$ can be computed in time $O((nL)^2(d|P| + d \log nL + |D|) + nL \sum_{i=1}^d |E_i|)$, where E_i is the set of pairs $(\alpha, \beta) : \beta \succ \alpha$ for which message $M_i(\alpha, \beta)$ was correctly computed in step 3 of Algorithm 2.

Proof. First let us show that there is a simple scaling procedure [9] that makes all edge weights nonnegative. Let us add a new vertex s (source) to $G = (\Pi, \succ)$ and connect this source with all vertices that have no incoming edges. We assign zero weights to new added edges. Since our graph is acyclic, then single source shortest path algorithm will take only $O((nL)^2)$ time [5]. This way we can compute values $r(\alpha), \alpha \in \Pi$ that are equal to the distance from s to α in the new graph. Now it can be checked that

$$r(\alpha) - r(\beta) + c(\alpha, \beta) \geq 0, \beta \succ \alpha \quad (20)$$

Therefore, we can define new weights by the following rescaling formula: $c'(\alpha, \beta) = r(\alpha) - r(\beta) + c(\alpha, \beta)$ and for every path from α to β in the graph its new length l' will depend on the old one l by formula: $l' = r(\alpha) - r(\beta) + l$. Therefore, rescaled graph will have the same shortest paths as the initial one.

Now for this rescaled version we can apply standard algorithms for all-pairs shortest path problem with nonnegative edge weights. We suggest an algorithm of Karger-Koller-Phillips [10] (or, of Demetrescu-Italiano[6]). This algorithm has complexity $O(|E^*||V| + |V|^2 \log |V|)$, where V is a set of vertices, E is a set of edges, and E^* is the set of edges $(u, v) \in E$ for which one of optimal paths from u to v is equal to $\{(u, v)\}$. Under a non-critical assumption we have $E^* = E_i$ for the graph (V, E) in step 4 of the i -th iteration of the algorithm.² Therefore, the total complexity of Algorithm 2 will be $O((nL)^2(d|P| + d \log nL + |D|) + nL \sum_{i=1}^d |E_i|)$. \square

The complexity is usually dominated by the term $nL \sum_{i=1}^d |E_i|$: in the worst case we have $|E_i| = O((nL)^2)$, and thus the overall worst-case complexity is cubic with respect to nL (as of the algorithm for general context-free grammars). In practice, however, $|E_i|$ can be smaller (reaching $|E_i| = O(nL)$ in the best case), and so the empirical runtime of the algorithm can be better than $O((nL)^3)$. Experiments on a synthetic data given in section 6 confirm this.

4.3 Linear time algorithm for $d = 1$

In this section we focus on the case $d = 1$. First, we show that just a simplification of Algorithm 2 gives the following.

Theorem 9 *For $d = 1$, $M = \min_x F(x)$ can be computed in time $O((nL)^2(|P| + |D|))$.*

Proof. The specificity of this case is determined by the fact that we compute shortest paths only for one graph and in the end of the algorithm in step 7 we need only shortest paths that starts from vertex ε_0 . Therefore, instead of using all-pairs shortest path algorithm we can use only single source version of it. Since there is $O(|V| + |E|)$ single source shortest path algorithm for general acyclic graphs [5], overall complexity of Algorithm 2 is $O((nL)^2(|P| + |D|))$. \square

²To achieve this equality, we can subtract a sufficiently small value $\varepsilon > 0$ from initial edge weights; shortest paths of the graph with new weights will also be shortest for the graph with old weights.

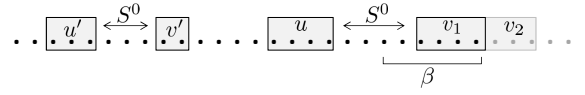


Figure 5: A state $g = uS^0v1.v2, v2 \neq \varepsilon$. At a current stop at j_β we have already read a sequence $u \dots v_1$ and expect v_2 to come. All other interacting pairs (such as pair (u', v') shown in the figure) can be located only before u .

In the previous algorithm we did not change data structures, though made computations with them more efficient. This way we cannot radically improve complexity, since only reading the input would take time quadratic with respect to nL .

We now describe an alternative approach which scales linearly with nL . The idea is to compute another set of messages $S(g, \beta)$, where $\beta \in \Pi$ and g is a state described by a rule $r = S^1 \rightarrow uS^0v$ together with index pointing to a particular position inside this rule. More precisely, g can be one of the following: (a) $g = u_1.u_2S^0v$ with $u = u_1u_2$, (b) $g = uS^0v_1.v_2$ with $v = v_1v_2$, or (c) $g = u\dot{S}^0v$. The dot indicates the position inside r , and corresponds to the end of pattern β . The state $g = u\dot{S}^0v$ designates that the position is strictly inside the word between u and v . Note that u, v are words, not patterns (i.e. they are not associated with any interval).

Message $S(g, \beta)$ is defined as the minimum of the functional that includes costs of both patterns and rules over all partial assignments $x = x_{1:j_\beta}$ under two constraints: (i) $x = *\beta$ and $x \neq *\gamma, (\beta, \gamma) \in \Pi_{j_\beta}$; (ii) x can be extended to some assignment xy that can be derived using rule r together with all other rules counted for x , and the dot in $x.y$ would correspond to the dot in g (see Fig. 5). The cost of rule r is counted only if $g = uS^0v$.³

Theorem 10 *All messages $S(g, \beta)$, and therefore $M = \min_x F(x)$, can be computed in time $O(|P|nL(l_{\min} \min(|D|, \log l_{\min}) + |P|))$.*

Proof. We compute these messages in the order of increasing j_β . Let us define for a given state g a set P_g of possible states one step back (here and below $a \in D$): $P_{u_1a.u_2S^0v} = \{u_1.au_2S^0v\}$; $P_{uS^0v_1a.v_2} = \{uS^0v_1.av_2\}$, if $v_2 \neq \varepsilon$; $P_{uaS^0.v} = \{ua\dot{S}^0v, ua.S^0v, u.aS^0v\}$; $P_{u\dot{S}^0v} = \{u\dot{S}^0v, u.S^0v\}$.

If $g \neq .uS^0v$, it is easy to see that $S(g, \beta)$ can be calculated as a function of $S(g', \gamma), g' \in P_g, \gamma \in \Pi_{j_\beta-1}$ in the same way as it is done in algorithm for minimizing PBs [16]. Recall, that the average price of such computation is $O(\min(|D|, \log l_{\min}))$ where $l_{\min} = \min_{w \in \Lambda} |w|$ which leads to overall complexity of such computation $O(|P|l_{\min}nL \cdot \min(|D|, \log l_{\min}))$. The only specifics is a case of $g = uS^0va$, for which we also have to add a weight of rule $S^1 \rightarrow uS^0v$ to an expression when we calculate

³Note that similar states g are also used in Earley parser [8]. Thus, our algorithm can be viewed as an extension of Earley parser to GPBs. Unfortunately, for general grammars such extension would give algorithms whose complexity is non-polynomial in $|R|$.

$S(g, \beta)$ based on the assumption that the previous state was $uS^0v.a$.

For $g = .uS^0v$, first we compute $S(g, \beta)$ as a function of $S(g, \gamma)$, $\gamma \in \Pi_{j_\beta-1}$, plus we should take into consideration that $S(g, \beta) \leftarrow \min_{(u', v') \in P} S(u'S^0v', \beta)$.

After adding complexities for all cases we obtain overall complexity of $O(|P| \cdot nL(l_{\min} \min(|D|, \log l_{\min}) + |P|))$. \square

4.4 Generalizations for rules weights

So far we assumed for simplicity that the weights $\nu(r)$ of rules $r \in R$ do not depend on the interval $[i, j]$ for which this rule is applied. In practice, however, such dependence is desirable, since e.g. the input substring $z_{i:j}$ may vary for different intervals. We can incorporate this dependence by introducing weight $\nu(r, i, j)$ of a rule $r \in R$ given that we derive subword $x_{i:j}$ from its left-side nonterminal. It is straightforward to modify algorithms to this case (without affecting the complexity): when computing $M_A(\alpha, \beta)$, we simply need to use $\nu(r, j_\alpha + 1, j_\beta)$ instead of $\nu(r)$. The only exception is the algorithm in Theorem 10, since it works with different messages. In this case we can show the following.

Theorem 11 *Suppose that weights of rules $r \in R$ with intervals $[i, j]$ satisfy the property $\nu(r, i, j) = f(r, i) + g(r, j)$. Then $M = \min_x F(x)$ can be computed in time $O(|P|nL(l_{\min} \min(|D|, \log l_{\min}) + |P|))$.*

Proof. To prove this statement we only have to redefine messages $S(g, \beta)$ in a way that if $g = uS^0v_1.v_2$, $v = v_1v_2$ or $g = u\hat{S}^0v$ or $g = u.S^0v$, then the weight $f(S^1 \rightarrow uS^0v, i^*)$ should be present in $S(g, \beta)$, where i^* is an index from which u started. The calculation of messages is only slightly different from the previous and can be easily reconstructed. \square

5 Learning GPB model

We introduced a new family of probabilistic distributions that we call a grammatical pattern-based model. This distribution is defined on a pair of objects, i.e. $p(x, \lambda) \sim \exp\{-E_\theta(x, \lambda|z)\}$, where x stands for a labeling sequence and λ stands for a derivation of x according to some grammar Γ . Suppose that λ is a hidden variable and we need to learn the model. We showed that minimizing the energy over both x and λ is a tractable problem. Moreover, minimizing the energy over λ for a fixed x is equivalent to a least-weight parsing of x according to grammar Γ , which can be solved in $O(|R|n^3)$ time by the standard CYK algorithm[1]. Together, these two facts open a possibility to learn GPB models (under condition that we parameterize energy linearly with respect to weights to be learned) by the struct-SVM with hidden variables approach [19].

Learning the model with maximum likelihood approach by either gradient-based or EM-based methods [13] requires another kind of algorithms. First of all we need algorithms for computing expressions like

$\sum_{x, \lambda} \exp\{-E_\theta(x, \lambda|z)\}$ and $\sum_\lambda \exp\{-E_\theta(x, \lambda|z)\}$. It can be seen that our algorithm for general WCFG can be turned into an algorithm for computing the first sum. Indeed, if in the definition of messages and f, ϕ -expressions we replace “min” with “ \sum ” and “+” with multiplication (and accordingly, “−” in the algorithm is turned into division and new weights of patterns and rules are defined as $e^{-\text{old weight}}$) we obtain a valid algorithm for computing such sums. Moreover, now we can compute the matrix product in step 3 of the algorithm using some fast matrix multiplication algorithm [4], which leads to complexity $O(|R|(nL)^{2.376})$. Also, sums of the second type can be computed in time $O(|R|n^{2.376})$.

Such sums allow computing in polynomial time marginals required by gradient-based or EM-based methods of learning, e.g. by running the algorithm independently for each marginal with appropriately modified costs. We conjecture, however, that the marginals can be computed more efficiently by a single computation, similar to [18, 16]. This is left as a future work.

6 Experiments and discussion

To support the claim the the runtime of the algorithm for interaction grammars can be better than $O(n^3)$ (for a fixed set of patterns and interacting pairs), we present some computational results on a synthetic data. As a subroutine for solving all-pairs shortest path problem in step 4 of Algorithm 2 we used a code based on [7] that we took from <http://www.dis.uniroma1.it/~demetres/experim/dsp/>⁴.

In GPB potential we defined $D = \{0, 1\}$ and $\Lambda = D^4$, i.e. $|\Lambda| = 16$. An interaction grammar Γ of depth 2 contained only one “interaction” rule $r = S \rightarrow 11S11$. For each pattern $\alpha = [i_\alpha, j_\alpha, c_\alpha]$ its weight c_α was taken as a uniformly distributed random value from interval $[0, 1]$. A weight $\nu(r, i, j)$ was taken as a uniformly distributed random value from interval $[0, C]$, where C is a parameter that we varied from 0.0 to 10 with step 0.1. We introduced a parameter C to consider cases when pattern-based part is dominant in GPB potential ($C = 0.0$) and visa versa ($C = 10.0$). The length of variable chain was varied from 10 to 350 with step 10.

The dependence of minimization time T on chain length n for different values of C is shown in Fig. 6. For all values of C , in the interval $[100, 350]$ minimization time starts to depend on n as n^x with a power x between 2.2 and 2.3. Our experiments also showed that the dependence does not change whether we make subtractions of small values from edge weights that we described. A very similar dependence of T on n (as $n^{2.3}$) was obtained for other choices of “interaction” rule r , like e.g., $S \rightarrow 0S1$ or $S \rightarrow 1S1$.

⁴Note that for Demetrescu-Italiano algorithm the complexity bound of $O(|E^*||V| + |V|^2 \log |V|)$ is true only if all shortest paths are unique. This condition we can satisfy by additional subtracting of some random value from interval $[0, \varepsilon']$ for sufficiently small ε' from each edge weight.

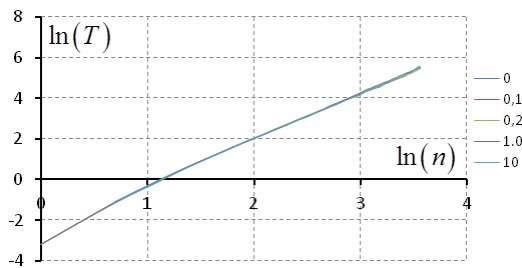


Figure 6: The dependence of minimization time T on chain length n for different values of C .

7 Conclusions

The GPB model can be viewed as a natural combination of a local PB and a more global WCFG models: combining different constraints by taking a sum of energy terms is a standard approach in the CRF literature. We showed that various inference tasks in GPBs can be solved in polynomial time.

The complexity of our general-purpose algorithm is rather high, and it can be prohibitively slow for some applications. However, we showed there exist faster techniques in some special case, namely interaction grammars of a fixed depth. This suggests that there may be other classes of GPBs with better complexity (such as $LR(k)$ grammars). We hope that our paper will stimulate the search for such classes.

References

- [1] Alfred V. Aho and John E. Hopcroft. *The Design and Analysis of Computer Algorithms*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition, 1974.
- [2] Yehoshua Bar-Hillel, M. Perles, and E. Shamir. On formal properties of simple phrase structure grammars. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 14:143–172, 1961. Reprinted in Y. Bar-Hillel. (1964). *Language and Information: Selected Essays on their Theory and Application*, Addison-Wesley 1964, 116–150.
- [3] José-Miguel Benedí and Joan-Andreu Sánchez. Combination of N-grams and stochastic context-free grammars for language modeling. In *COLING*, pages 55–61, 2000.
- [4] D. Coppersmith and S. Winograd. Matrix multiplication via arithmetic progressions. In *Proceedings of the Nineteenth Annual ACM Symposium on Theory of Computing*, STOC ’87, pages 1–6, 1987.
- [5] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. The MIT Press, 2 edition, 2001.
- [6] C. Demetrescu and G. Italiano. A new approach to dynamic all pairs shortest paths. *J. ACM*, 51(6):968–992, 2004.
- [7] Camil Demetrescu, Stefano Emiliozzi, and Giuseppe F. Italiano. Experimental analysis of dynamic all pairs shortest path algorithms. In *SODA*, pages 369–378, 2004.
- [8] Jay Earley. An efficient context-free parsing algorithm. *Commun. ACM*, 13(2):94–102, February 1970.
- [9] Andrew V. Goldberg. Scaling algorithms for the shortest paths problem. In *SODA*, pages 222–231, 1993.
- [10] D.R. Karger, D. Koller, and S. J. Phillips. Finding the hidden path: time bounds for all-pairs shortest paths. *SIAM Journal on Computing*, 22(6):1199–1217, 1993. Full version of paper in FOCS ’91.
- [11] George Katsirelos, Nina Narodytska, and Toby Walsh. The weighted grammar constraint. *Annals of Operations Research*, 184:179–207, 2011.
- [12] Viet Cuong Nguyen, Nan Ye, Wee Sun Lee, and Hai Leong Chieu. Semi-Markov conditional random field with high-order features. In *ICML 2011 Structured Sparsity: Learning and Inference Workshop*, 2011.
- [13] S. Nowozin and C. H. Lampert. Structured learning and prediction in computer vision. *Foundations and Trends in Computer Graphics and Vision*, 6:185–365, 2011.
- [14] Xian Qian, Xiaoqian Jiang, Qi Zhang, Xuanjing Huang, and Lide Wu. Sparse higher order conditional random fields for improved sequence labeling. In *ICML*, 2009.
- [15] Ismael Salvador and José-Miguel Benedí. RNA modeling by combining stochastic context-free grammars and n-gram models. *International Journal of Pattern Recognition and Artificial Intelligence*, 16:309–315, 2002.
- [16] R. Takhanov and V. Kolmogorov. Inference algorithms for pattern-based CRFs on sequence data. In *ICML*, 2013.
- [17] Yeyi Wang, Milind Mahajan, and Xuedong Huang. A unified context-free grammar and n-gram model for spoken language processing. In *International Conference of Acoustics, Speech, and Signal Processing*, pages 1639–1642, 2000.
- [18] Nan Ye, Wee Sun Lee, Hai Leong Chieu, and Dan Wu. Conditional random fields with high-order features for sequence labeling. In *NIPS*, 2009.
- [19] Chun-Nam John Yu and Thorsten Joachims. Learning structural SVMs with latent variables. In *ICML*, pages 1169–1176, 2009.

- [20] U. Zwick. All pairs shortest paths using bridging sets and rectangular matrix multiplication. *J. ACM*, 49(3):289317, 2002.